
# Regression

RAD Lab Machine Learning Workshop

Kurt Miller

08/23/07

Adapted from slides by Romain Thibaux

# Outline

- Ordinary Least Squares Regression
  - Online version
  - Normal equations
  - Probabilistic interpretation
- Overfitting and Regularization
- Overview of additional topics
  - $L_1$ Regression
  - Quantile Regression
  - Kernel Regression and LWR
  - Spline Regression

# Outline

- Ordinary Least Squares Regression
  - Online version
  - Normal equations
  - Probabilistic interpretation
- Overfitting and Regularization
- Overview of additional topics
  - $L_1$ Regression
  - Quantile Regression
  - Kernel Regression and LWR
  - Spline Regression

# Regression vs. Classification:

# Classification

$$X \rightarrow Y$$

Anything:

- continuous ($\Re$, $\Re^d$, …)
- discrete ({0,1}, {1,…k}, …)
- structured (tree, string, …)
- …

- discrete:
  - {0,1} *binary*
  - {1,…k} *multi-class*
  - tree, etc. *structured*

# Regression vs. Classification:

# Classification

$$X \rightarrow Y$$

Anything:

- continuous ($\Re$, $\Re^d$, …)
- discrete ({0,1}, {1,…k}, …)
- structured (tree, string, …)
- …

Kernel trick

Perceptron
Logistic Regression
Support Vector Machine

Decision Tree
Random Forest

# Regression vs. Classification:

# Regression

$$X \rightarrow Y$$

Anything:

- continuous ($\Re$, $\Re^d$, …)
- discrete ({0,1}, {1,…k}, …)
- structured (tree, string, …)
- …

- continuous:
  - $\Re$, $\Re^d$

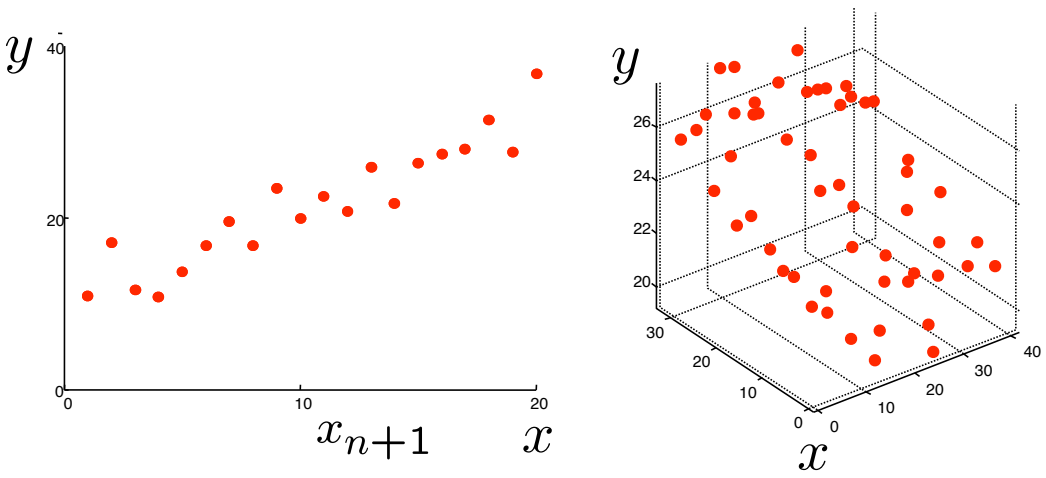
# Examples

- Voltage $\rightarrow$ Temperature
- Processes, memory $\rightarrow$ Power consumption
- Protein structure $\rightarrow$ Energy
- Robot arm controls $\rightarrow$ Torque at effector
- Location, industry, past losses $\rightarrow$ Premium

# Linear regression

Given examples $(x_i, y_i)_{i=1...n}$

Predict $y_{n+1}$ given a new point $x_{n+1}$

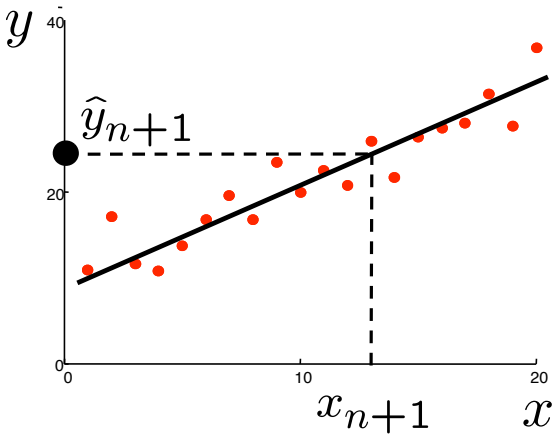


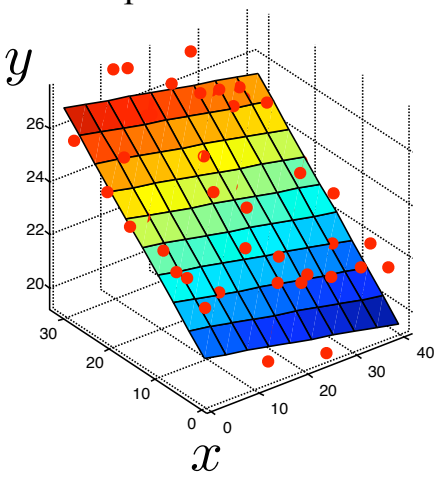
# Linear regression

We wish to estimate $\hat{y}$ by a linear function of our data $x$:

$$\begin{aligned}\hat{y}_{n+1} &= w_0 + w_1 x_{n+1,1} + w_2 x_{n+1,2} \\ &= w^\top x_{n+1}\end{aligned}$$

where $w$ is a parameter to be estimated and we have used the standard convention of letting the first component of $x$ be 1.

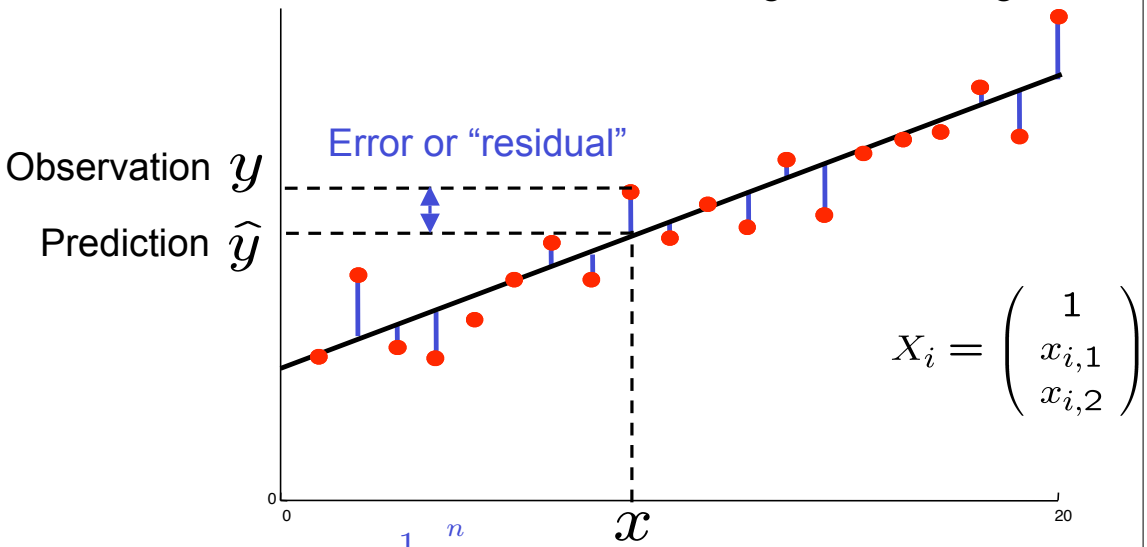



# LMS Algorithm

(Least Mean Squares)

In order to clarify what we mean by a good choice of $w$, we will define a cost function for how well we are doing on the training data:




$$\text{Cost} = \frac{1}{2}\sum_{i=1}^{n}(w^\top x_i - y_i)^2$$

# LMS Algorithm
(Least Mean Squares)

The best choice of $w$ is the one that minimizes our cost function

$$E = \frac{1}{2}\sum_{i=1}^{n}(w^\top x_i - y_i)^2 = \sum_{i=1}^{n} E_i$$

In order to optimize this equation, we use standard gradient descent

$$w := w - \alpha \frac{\partial}{\partial w} E$$

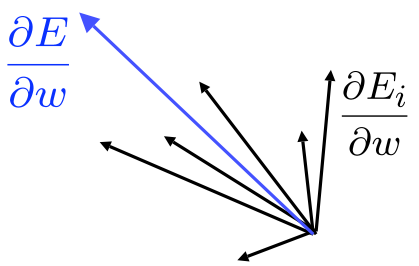
where

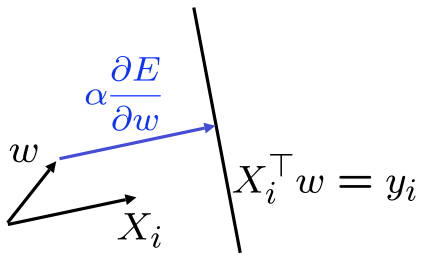
$$\frac{\partial}{\partial w}E = \sum_{i=1}^{n}\frac{\partial}{\partial w}E_i \qquad \text{and} \qquad \begin{aligned}\frac{\partial}{\partial w}E_i &= \frac{1}{2}\frac{\partial}{\partial w}(w^\top x_i - y_i)^2 \\ &= (w^\top x_i - y_i)x_i\end{aligned}$$

# LMS Algorithm
(Least Mean Squares)

The LMS algorithm is an online method that performs the following update for each new data point

$$\begin{aligned} w^{t+1} &:= w^t - \alpha \frac{\partial}{\partial w} E_i \\ &= w^t + \alpha(y_i - x_i^\top w)x_i \end{aligned}$$



# LMS, Logistic regression, and Perceptron updates

- LMS

$$w^{t+1} := w^t + \alpha(y_i - x_i^\top w)x_i$$

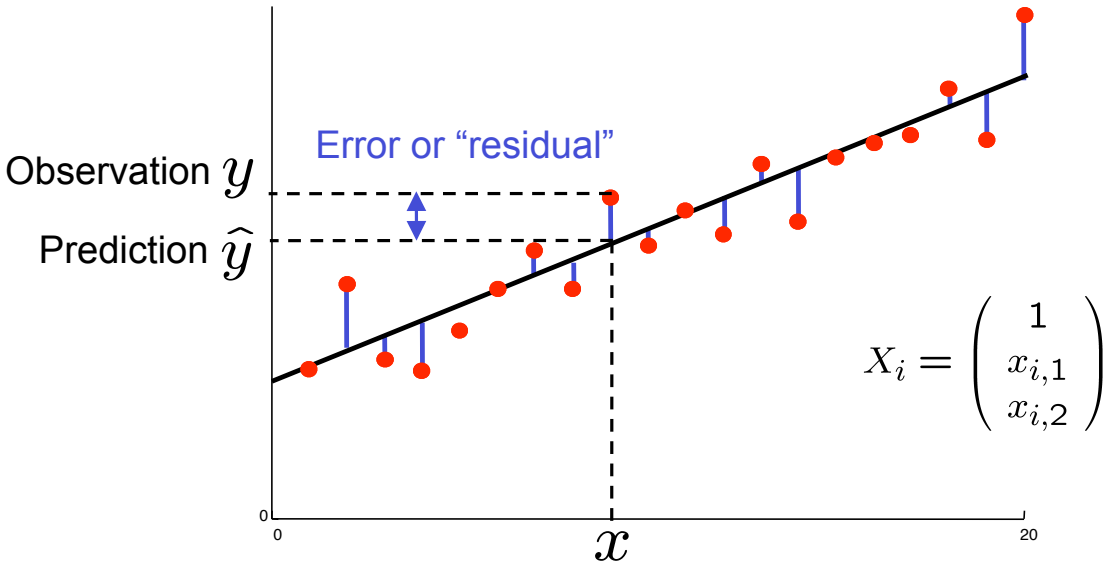
- Logistic Regression

$$w^{t+1} := w^t + \alpha(y_i - f_w(x_i))x_i$$

- Perceptron

$$w^{t+1} := w^t + \alpha(y_i - f_w(x_i))x_i$$

# Ordinary Least Squares (OLS)




$$\text{Cost} = \frac{1}{2}\sum_{i=1}^{n}(w^\top x_i - y_i)^2$$

# Minimize the sum squared error

$$
\begin{aligned}
E &= \frac{1}{2}\sum_{i=1}^{n}(w^\top x_i - y_i)^2 \\
&= \frac{1}{2}\|Xw - y\|_2^2 \\
&= \frac{1}{2}\left(w^\top X^\top X w - 2y^\top X w + y^\top y\right)
\end{aligned}
$$

$$
\frac{\partial}{\partial w}E = X^\top X w - X^\top y
$$

$$
X = \begin{pmatrix} -x_1^\top - \\ -x_2^\top - \\ \cdots \end{pmatrix}
$$

n

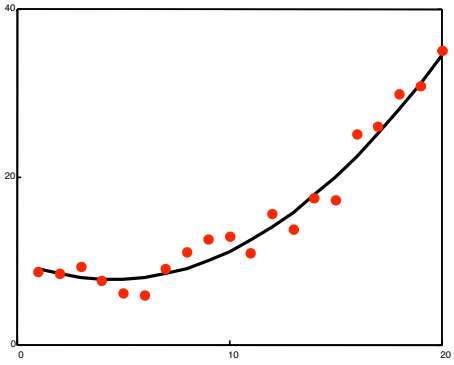
d

Setting the derivative equal to zero gives us the *Normal Equation*

$$
\begin{aligned}
X^\top X w &= X^\top y \\
w &= (X^\top X)^{-1} X^\top y
\end{aligned}
$$

# Beyond lines and planes

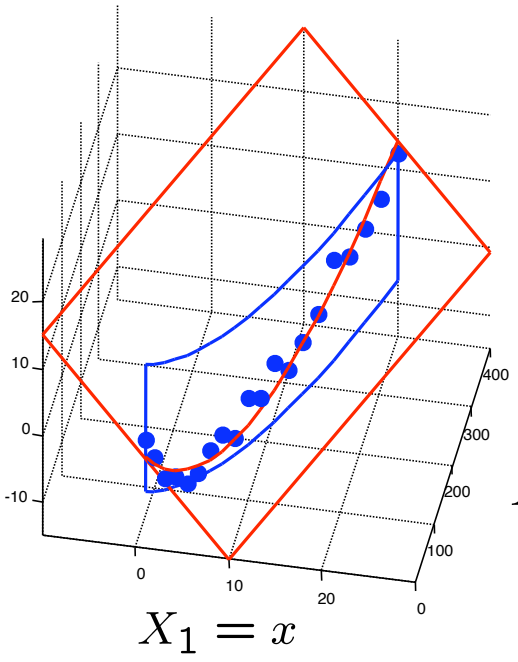
$$\hat{y}_i = w_0 + w_1 x_i + w_2 x_i^2$$




still linear in $w$

everything is the same with $X_i = \begin{pmatrix} 1 \\ x_i \\ x_i^2 \end{pmatrix}$

# Geometric interpretation




$X_1 = x$

$X_2 = x^2$

$$\hat{y} = w_0 + w_1 x + w_2 x^2$$

[Matlab demo]

# Ordinary Least Squares [summary]

Given examples $(x_i, y_i)_{i=1...n}$

Let $X_i^\top = (f_1(x_i) \quad f_2(x_i) \quad \dots \quad f_d(x_i))$

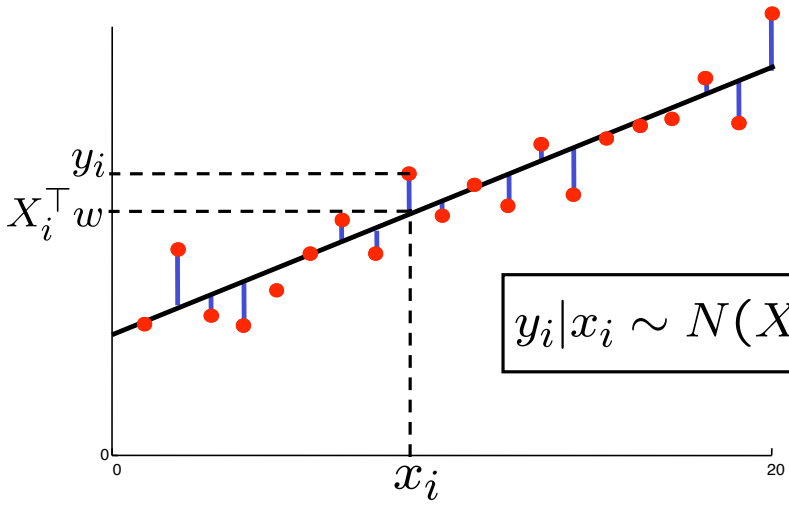
For example $X_i^\top = \begin{pmatrix} 1 & x_{i,1} & x_{i,2} & x_{i,1}^2 & x_{i,2}^2 & x_{i,1}x_{i,2} \end{pmatrix}$

Let $X = \begin{pmatrix} -X_1^\top- \\ -X_2^\top- \\ \dots \end{pmatrix}$ n (d) $\qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \end{pmatrix}$

Minimize $\|Xw - y\|_2^2$ by solving $\left(X^\top X\right) w = X^\top y$

Predict $\hat{y}_{n+1} = X_{n+1}^\top w$

# Probabilistic interpretation



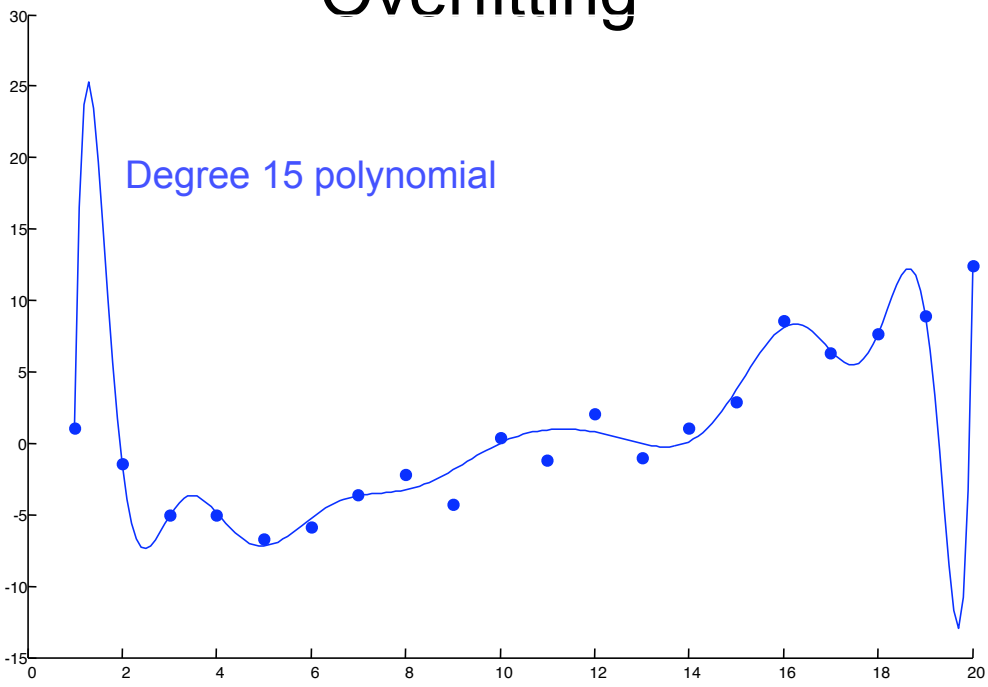

Likelihood $L = \prod_i \exp -\frac{1}{2\sigma^2}(X_i^\top w - y_i)^2 = \exp -\frac{1}{2\sigma^2}\sum_i (X_i^\top w - y_i)^2$

$\underset{w}{\text{argmax}}\, L = \underset{w}{\text{argmin}}\, E$

# Outline

- Ordinary Least Squares Regression
  - Online version
  - Normal equations
  - Probabilistic interpretation
- Overfitting and Regularization
- Overview of additional topics
  - $L_1$ Regression
  - Quantile Regression
  - Kernel Regression and LWR
  - Spline Regression

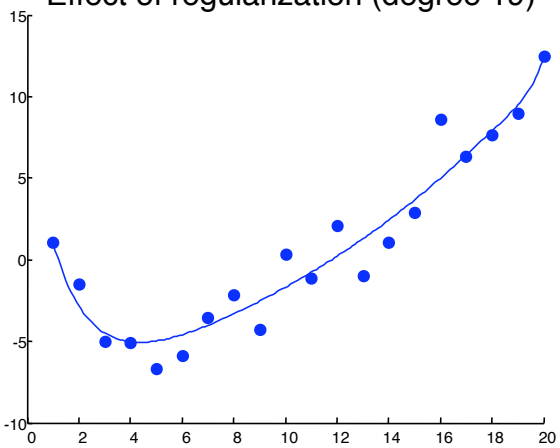
# Overfitting




[Matlab demo]

# Ridge Regression (Regularization)

Effect of regularization (degree 19)




$$A = X^\top X$$

$$b = X^\top y$$

with $\epsilon$ "small"

Minimize $\frac{1}{2}\|Xw - y\|_2^2 + \epsilon\|w\|_2^2$ by solving $(A + \epsilon I)w = b$

# Probabilistic interpretation

Likelihood $y_i|x_i \sim N(X_i^\top w, \sigma^2)$

Prior $w \sim N\left(0, \frac{\sigma^2}{\epsilon}\right)$

Posterior $P(w|x_1, \ldots x_n) = \frac{P(w, x_1, \ldots x_n)}{P(x_1, \ldots x_n)} \propto P(w, x_1, \ldots x_n)$

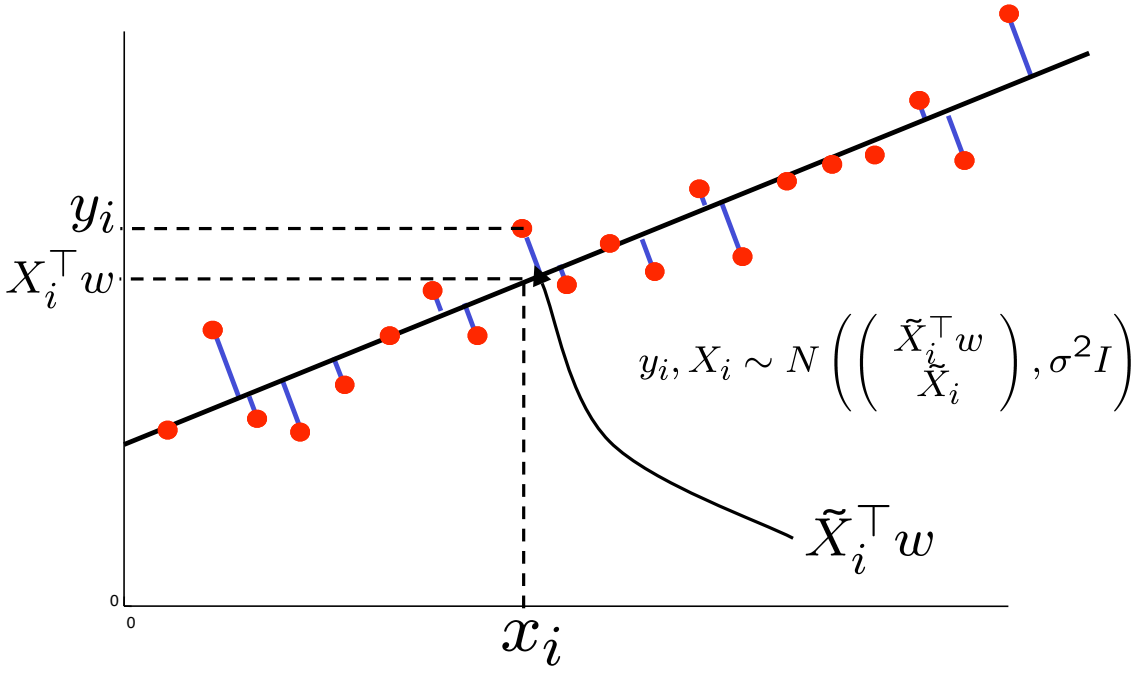
$$P(w, x_1, \ldots x_n) = \exp -\left\{\frac{\epsilon}{2\sigma^2}\|w\|_2^2\right\} \prod_i \exp -\frac{1}{2\sigma^2}(X_i^\top w - y_i)^2$$

$$= \exp -\frac{1}{2\sigma^2}\left[\epsilon\|w\|_2^2 + \sum_i (X_i^\top w - y_i)^2\right]$$

# Outline

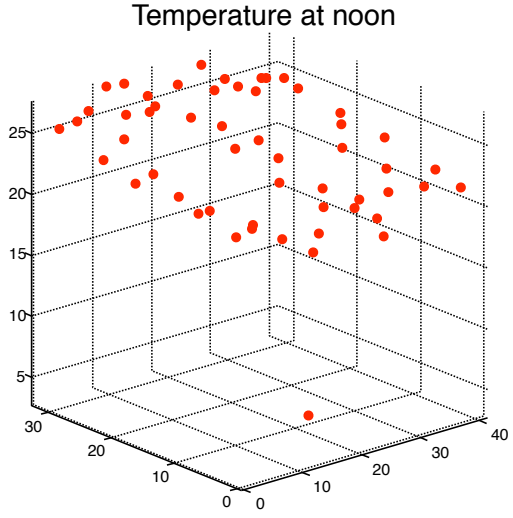
- Ordinary Least Squares Regression
  - Online version
  - Normal equations
  - Probabilistic interpretation
- Overfitting and Regularization
- Overview of additional topics
  - $L_1$ Regression
  - Quantile Regression
  - Kernel Regression and LWR
  - Spline Regression

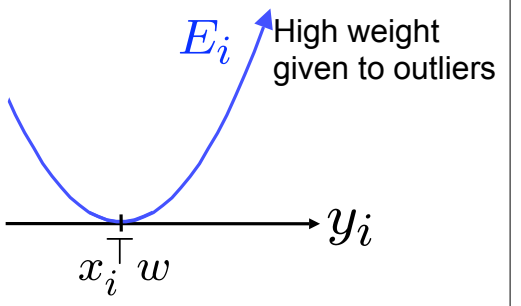
# Errors in Variables (Total Least Squares)

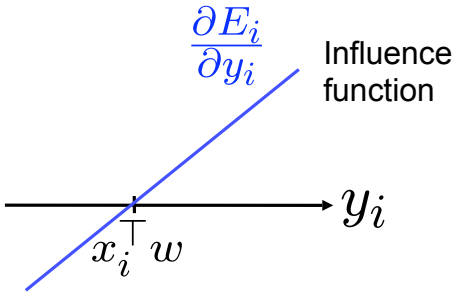



# Sensitivity to outliers

$$E = \sum_i (x_i^\top w - y_i)^2 = \sum_i E_i$$

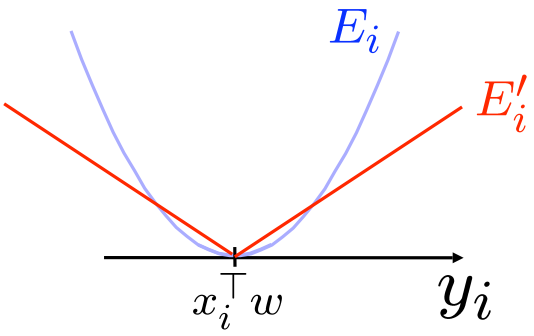


# $L_1$ Regression

$$E' = \sum_i |x_i^\top w - y_i|$$
$$= \sum_i E'_i$$

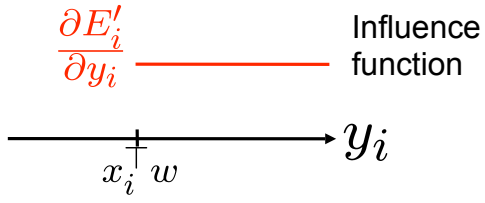



Linear program

$$\min_{w,c} \sum_i c_i$$

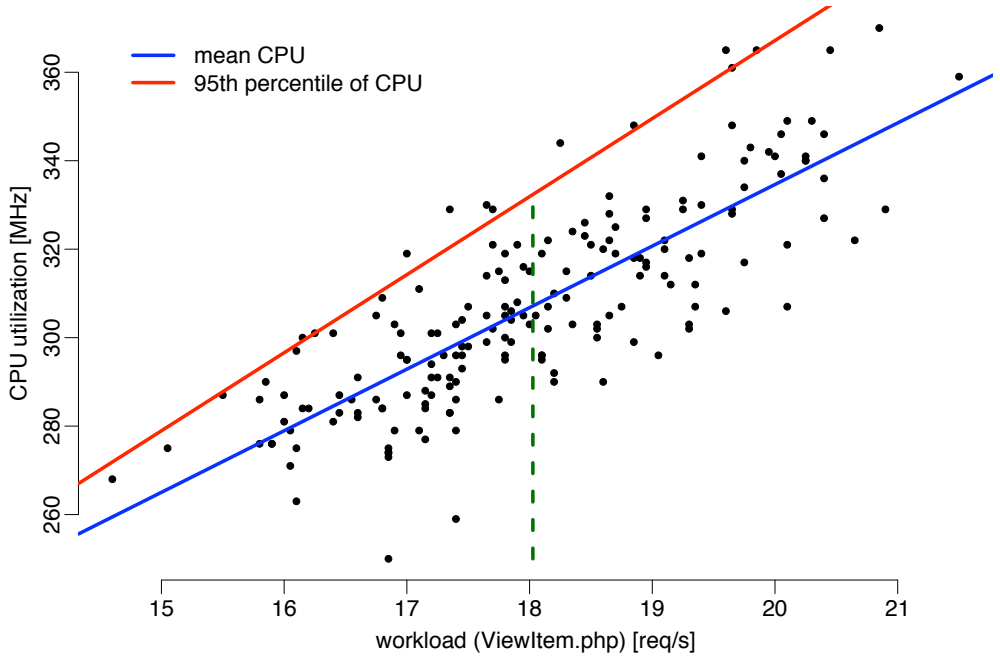
s.t. $x_i^\top w - y_i \le c_i \quad \forall i$

$y_i - x_i^\top w \le c_i \quad \forall i$




[Matlab demo]

# Quantile Regression




Slide courtesy of Peter Bodik

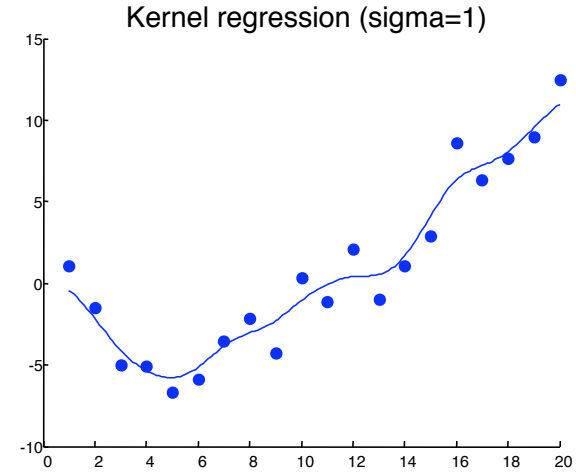
# Kernel Regression and Locally Weighted Linear Regression

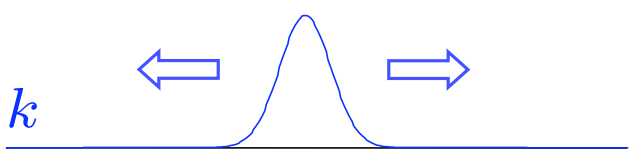
- **Kernel Regression:**
  Take a very very conservative function approximator called AVERAGING. Locally weight it.

- **Locally Weighted Linear Regression:**
  Take a conservative function approximator called LINEAR REGRESSION. Locally weight it.

Slide from Paul Viola 2003

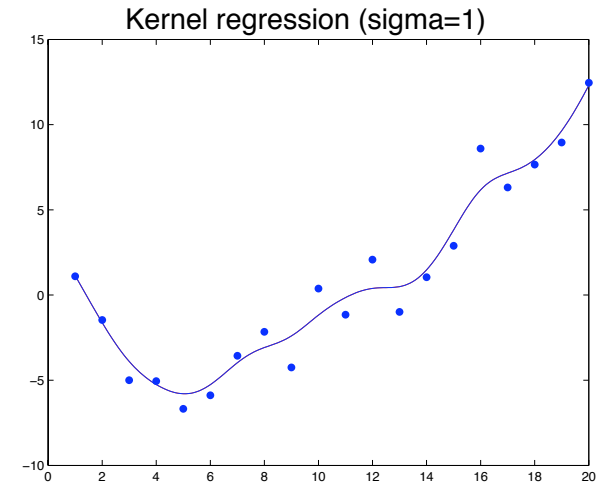
# Kernel Regression

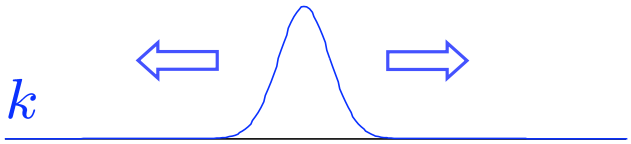



$$\hat{y}(x) = \frac{\sum_i y_i k(x_i - x)}{\sum_i k(x_i - x)}$$

# Locally Weighted Linear Regression (LWR)




OLS cost function:

$$E = \frac{1}{2}\sum_{i=1}^{n}(w^\top x_i - y_i)^2$$

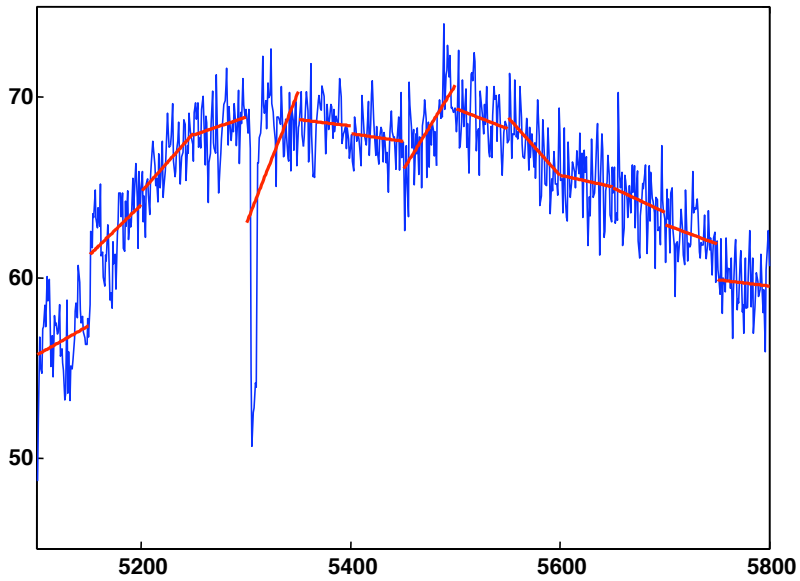
LWR cost function:

$$E' = \sum_{i=1}^{n} k(x_i - x)(w^\top x_i - y_i)^2$$

[Matlab demo]

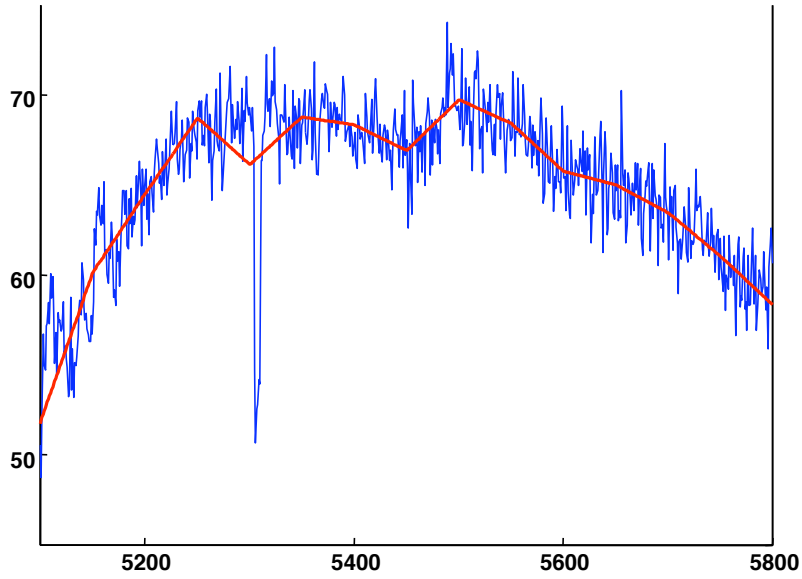
# Spline Regression

## Regression on each interval



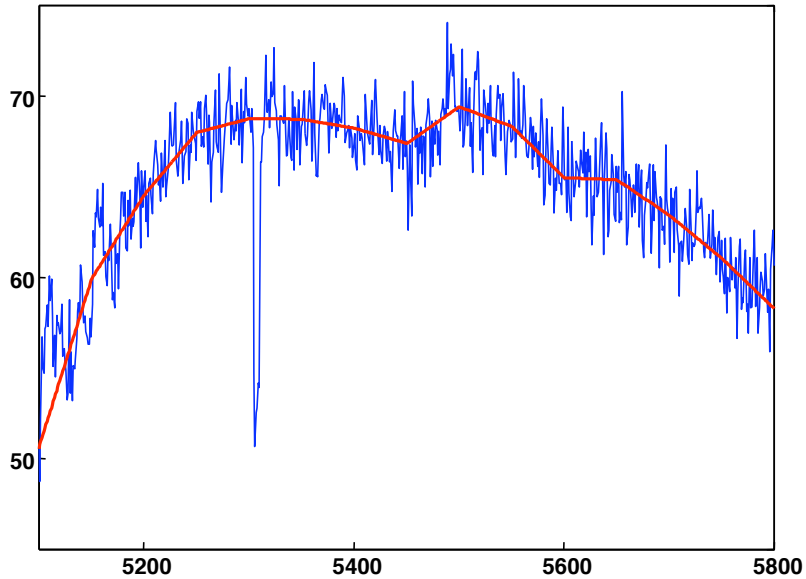
# Spline Regression

## With equality constraints




# Spline Regression

## With $L_1$ cost



# Further topics

- Generalized Linear Models
- Gaussian process regression
- Feature Selection [after the break]